Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 2, *pp.* 143–170

On the use of non-local prior densities in Bayesian hypothesis tests

Valen E. Johnson

M. D. Anderson Cancer Center, Houston, USA

and David Rossell

Institute for Research in Biomedicine, Barcelona, Spain

[Received March 2009. Revised October 2009]

Summary. We examine philosophical problems and sampling deficiencies that are associated with current Bayesian hypothesis testing methodology, paying particular attention to objective Bayes methodology. Because the prior densities that are used to define alternative hypotheses in many Bayesian tests assign non-negligible probability to regions of the parameter space that are consistent with null hypotheses, resulting tests provide *exponential* accumulation of evidence in favour of true alternative hypotheses, but only *sublinear* accumulation of evidence in favour of true null hypotheses. Thus, it is often impossible for such tests to provide strong evidence in favour of a true null hypothesis, even when moderately large sample sizes have been obtained. We review asymptotic convergence rates of Bayes factors in testing precise null hypotheses and propose two new classes of prior densities that ameliorate the imbalance in convergence rates that is inherited by most Bayesian tests. Using members of these classes, we obtain analytic expressions for Bayes factors in linear models and derive approximations to Bayes factors in large sample settings.

Keywords: Fractional Bayes factor; Intrinsic Bayes factor; Intrinsic prior; Inverse moment density function; Moment density function; Objective Bayes analysis

1. Introduction

Since the advent of Markov chain Monte Carlo algorithms in the early 1990s, applications of Bayesian methodology to problems of statistical estimation and testing have increased dramatically. Much of this activity has been premised on the use of objective Bayesian models, or Bayesian models that use vague (i.e. non-informative or disperse) prior distributions on model parameter spaces. Objective Bayesian methodology is now commonly used for parameter estimation and inference, but unresolved philosophical and technical issues continue to limit the application of Bayesian methodology in the conduct of hypothesis tests.

In parametric settings, if $\boldsymbol{\theta} \in \Theta$ denotes the parameter of interest, then classical hypothesis tests are usually posed as a test of two hypotheses,

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \Theta_0, \\ H_1 &: \boldsymbol{\theta} \in \Theta_1, \end{aligned} \tag{1}$$

Address for correspondence: Valen E. Johnson, Department of Biostatistics, Division of Quantitative Sciences, M. D. Anderson Cancer Center, Unit 1411, 1400 Pressler Street, Houston, TX 77030, USA.
E-mail: vejohnson@mdanderson.org

1369–7412/10/72143

where Θ_0 and Θ_1 are disjoint and $\Theta_0 \cup \Theta_1 = \Theta$. Testing in the Bayesian paradigm requires specifying prior distributions on $\boldsymbol{\theta}$ under each hypothesis, which means that expression (1) might also be written as

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \sim \pi_0(\boldsymbol{\theta}), \\ H_1 &: \boldsymbol{\theta} \sim \pi_1(\boldsymbol{\theta}), \end{aligned} \tag{2}$$

provided that $\pi_0(\boldsymbol{\theta})$ and $\pi_1(\boldsymbol{\theta})$ are 0 on Θ_1 and Θ_0 respectively. Violations of this provision form the central focus of this paper, i.e. most Bayesian hypothesis tests are defined with alternative prior densities π_1 that are positive on Θ_0. We refer to such priors as 'local alternative prior densities'. We also exploit the correspondence between expressions (1) and (2) to associate prior densities with statistical hypotheses, and refer, for example, to an alternative hypothesis H_1 defined with a local prior as a 'local alternative hypothesis'.

On a philosophical level, we object to local alternative priors on the grounds that they do not incorporate any notion of a minimally significant separation between the null and alternative hypotheses. An alternative hypothesis, by definition, should reflect a theory that is fundamentally different from the null hypothesis. Local alternative hypotheses do not. The alternative hypotheses that are proposed in this paper do, but by necessity they require the specification of a scale parameter that implicitly defines the meaning of a substantively important deviation from the null hypothesis.

An impediment to the widespread application of Bayesian methodology to parametric hypothesis tests has been the requirement to specify proper prior distributions on model parameters. Improper prior distributions cannot be used to define Bayes factors between competing hypotheses in the usual way, and the deleterious effects of vaguely specified prior distributions do not diminish even as sample sizes become large (e.g. Lindley (1957)).

Numerous solutions have been proposed to solve this problem. As early as 1939, Jeffreys proposed to resolve this difficulty by defining priors under the alternative hypothesis that were centred on the null hypothesis. In testing a point null hypothesis that a parameter α was 0, Jeffreys reasoned

> 'that the mere fact that it has been suggested that α is zero corresponds to some presumption that it is fairly small'

(Jeffreys (1998), page 251). Since Jeffreys's proposal, most published Bayesian testing procedures have been premised on the use of local alternative hypotheses. Kass and Raftery (1995) provided a review of these and related Bayesian testing procedures through the mid-1990s (see also Lahiri (2001) and Walker (2004)).

More recently, fractional Bayes factor (O'Hagan, 1995, 1997; Conigliani and O'Hagan, 2000; De Santis and Spezzaferri, 2001) and intrinsic Bayes factor (Berger and Pericchi, 1996a, 1998; Berger and Mortera, 1999; Pérez and Berger, 2002) methodologies have been used to define default prior densities under alternative hypotheses implicitly. When data are generated from a model that is consistent with the null hypothesis, fractional Bayes factors and certain intrinsic Bayes factors (e.g. arithmetic, geometric, expected arithmetic, expected geometric and median Bayes factors) typically produce alternative prior densities that are positive at parameter values that are consistent with the null hypothesis. As a consequence, these Bayes factors share properties of Bayes factors that are defined by using more traditional local alternative priors.

In many settings, limiting arguments can be used to define the prior densities on which intrinsic Bayes factors are based (Berger and Pericchi, 1996a, 1998; Moreno *et al.*, 1998; Bertolino *et al.*, 2000; Cano *et al.*, 2004; Moreno, 2005; Moreno and Girón, 2005; Casella and Moreno, 2006). Because the resulting intrinsic (alternative) priors are defined with respect to a specific null hypoth-

esis, they concentrate prior mass on parameters that are consistent with the null model. As a consequence, this methodology also results in the specification of local alternative prior densities.

Bayes factors that are obtained by using local alternative priors exhibit a disturbing large sample property. As the sample size n increases, they accumulate evidence much more rapidly in favour of true alternative models than in favour of true null models. For example, when testing a point null hypothesis regarding the value of a scalar parameter, the Bayes factor in favour of H_0 is $O_p(n^{1/2})$ when data are generated from H_0. Yet, when data are generated from H_1, the *logarithm* of the Bayes factor in favour of H_1 increases at a rate that is linear in n (e.g. Bahadur and Bickel (1967), Walker and Hjort (2001) and Walker (2004)). The accumulation of evidence under true null and true alternative hypotheses is thus highly asymmetric, even though this fact is not reflected in probability statements regarding the outcome of a Bayesian hypothesis test (e.g. Vlachos and Gelfand (2003)).

Standard frequentist reports of the outcome of a test reflect this asymmetry. The null hypothesis is not accepted; it is simply not rejected. Because there is no possibility that the null hypothesis will be accepted, the rate of accumulation of evidence in its favour is less problematic. In contrast, a Bayesian hypothesis test based on a local alternative prior density results in the calculation of the posterior probability that the null hypothesis is true, even when this probability cannot (by design) exceed what is often a very moderate threshold. Verdinelli and Wasserman (1996) and Rousseau (2007) partially addressed this issue by proposing non-local alternative priors of the form $\pi(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta}$ in a neighbourhood of Θ_0, but the framework that results from this approach lacks flexibility in specifying the rate at which $\pi(\boldsymbol{\theta})$ approaches 0 near Θ_0, and it provides no mechanism for rejecting H_0 for values of $\boldsymbol{\theta}$ outside but near Θ_0.

In this paper, we describe two new classes of prior densities that more equitably balance the rates of convergence of Bayes factors in favour of true null and true alternative hypotheses. Prior densities from these classes offer a compromise between the use of vague proper priors, which can lead to nearly certain acceptance of the null hypothesis (Jeffreys, 1998; Lindley, 1957), and the use of local alternative priors, which restrict the accumulation of evidence in favour of the null hypothesis. These prior densities rely on a single parameter to determine the scale for deviations between the null and alternative hypotheses. Judicious selection of this parameter can increase the weight of evidence that is collected in favour of both true null and true alternative hypotheses. Our presentation focuses on the case of point null hypotheses (which comprise a vast majority of the null hypotheses that are tested in the scientific literature), although we briefly consider the extension of our methods to composite null and alternative hypotheses in the final discussion.

The remainder of this paper is organized as follows. In Section 2, we review the asymptotic properties of Bayes factors based on local alternative priors in regular parametric models. In Section 3, we propose and study the convergence properties of two families of non-local alternative prior densities: *moment prior densities* and *inverse moment prior densities*. In Section 4, we use members from these classes to obtain Bayes factors against precise null hypotheses in linear model settings. In particular, we obtain closed form expressions for Bayes factors based on moment prior densities that are defined by using Zellner's g-prior (Zellner, 1986), and we demonstrate that Bayes factors defined by using moment priors based on multivariate t-densities and inverse moment priors can be expressed as the expectation of a function of two gamma random variables. These results extend naturally to the definition of simple approximations to Bayes factors in large sample settings. We provide examples to illustrate the use of non-local alternative models in Section 5, and concluding remarks appear in Section 6.

The computer code that was used to produce the results in Table 2 can be obtained from `http://www.blackwellpublishing.com/rss`.

2. Local alternative prior densities

We begin by examining the large sample properties of Bayes factors defined by using local alternative prior densities. For this paper, we restrict our attention to parametric models that satisfy the regularity conditions that were used by Walker (1969) to establish asymptotic normality of the posterior distribution. These conditions are stated in Appendix A.

Following Walker's regularity conditions, let $X_1, \ldots, X_n$ denote a random sample from a distribution that has density function $f(x|\boldsymbol{\theta})$ with respect to a σ-finite measure μ, and suppose that $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^d$. Let

$$p_n(\mathbf{X}^{(n)}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta})$$

denote the joint sampling density of the data, let $L_n(\boldsymbol{\theta}) = \log\{p_n(\mathbf{X}^{(n)}|\boldsymbol{\theta})\}$ denote the log-likelihood function, and let $\hat{\boldsymbol{\theta}}_n$ denote a maximum likelihood estimate of $\boldsymbol{\theta}$. Let $\pi_0(\boldsymbol{\theta})$ denote either a continuous density function defined with respect to Lebesgue measure, or a unit mass concentrated on a point $\boldsymbol{\theta}_0 \in \Theta$. The density function π_1 is assumed to be continuous with respect to Lebesgue measure. If π_j, $j = 0, 1$, is continuous, define

$$m_j(\mathbf{X}^{(n)}) = \int_{\Theta} p_n(\mathbf{X}^{(n)}|\boldsymbol{\theta})\, \pi_j(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

to be the marginal density of the data under prior density π_j. The marginal density of the data for a specified value of $\boldsymbol{\theta}_0$ is simply the sampling density evaluated at $\boldsymbol{\theta}_0$. The Bayes factor based on a sample size n is defined as

$$\mathrm{BF}_n(1|0) = \frac{m_1(\mathbf{X}^{(n)})}{m_0(\mathbf{X}^{(n)})}.$$

The density $\pi_1(\boldsymbol{\theta})$ that is used to define H_1 in expression (2) is called a *local alternative prior density* (or, more informally, a *local alternative hypothesis*) if

$$\pi_1(\boldsymbol{\theta}) > \varepsilon \qquad \text{for all } \boldsymbol{\theta} \in \Theta_0. \tag{3}$$

Local alternative prior densities are most commonly used to test null hypotheses in which the dimension of Θ_0 is less than d, the common dimension of Θ and Θ_1. Examples of such tests include standard tests of point null hypotheses, as well as the tests that are implicit to variable selection and graphical model selection problems.

In contrast, if for every $\varepsilon > 0$ there is $\zeta > 0$ such that

$$\pi_1(\boldsymbol{\theta}) < \varepsilon \qquad \text{for all } \boldsymbol{\theta} \in \Theta : \inf_{\boldsymbol{\theta}_0 \in \Theta_0} |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \zeta, \tag{4}$$

then we define π_1 to be a *non-local alternative prior density*. Non-local alternative priors are commonly used in tests for which the dimensions of Θ_0 and Θ_1 both equal d. For example, Moreno (2005) described intrinsic priors for testing hypotheses of the form $H_0 : \theta \leqslant \theta_0$ *versus* $H_1 : \theta \geqslant \theta_0$. The densities that are described in Section 3 provide non-local alternatives that can be applied to tests in which the dimension of Θ_0 is less than the dimension of Θ_1.

Throughout the remainder of the paper we assume that $\pi_0(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta_0$, that $\pi_0(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Theta_1$ and that $\pi_1(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta_1$. For simplicity, in the remainder of this section we also restrict attention to scalar-valued parameters; we discuss extensions to vector-valued parameters in Section 3.3.

We now examine the convergence properties of tests based on local alternative priors when data are generated from both the null and the alternative models.

2.1. Data generated from the null model

Suppose that $H_0 : \theta = \theta_0, \theta_0 \in R$, is a point null hypothesis, and assume that data are generated from the null model. Then the marginal density of the data under the null hypothesis is

$$\begin{aligned} m_0(\mathbf{X}^{(n)}) &= p_n(\mathbf{X}^{(n)}|\theta_0) \\ &= \exp\{L_n(\theta_0)\}. \end{aligned} \tag{5}$$

From Walker (1969), if we assume that $\pi_1(\theta)$ is a local alternative prior density, it follows that $m_1(\mathbf{X}^{(n)})$ satisfies

$$\frac{m_1(\mathbf{X}^{(n)})}{\sigma_n \; p_n(\mathbf{X}^{(n)}|\hat{\theta}_n)} \xrightarrow{\mathrm{p}} \pi_1(\theta_0)\sqrt{(2\pi)}, \tag{6}$$

where $\sigma_n^2 = \{-L''(\hat{\theta}_n)\}^{-1}$. Because

$$L_n(\theta_0) - L_n(\hat{\theta}_n) = O_p(1)$$

and

$$\sigma_n = O_p(n^{-1/2}),$$

the Bayes factor in favour of the alternative hypothesis when the null hypothesis is true satisfies

$$\mathrm{BF}_n(1|0) = \frac{m_1(\mathbf{X}^{(n)})}{m_0(\mathbf{X}^{(n)})} = O_p(n^{-1/2}). \tag{7}$$

An example is presented in Section 5.1 for which $\mathrm{BF}_n(1|0) = \exp(B)(n+a)^{-1/2}$, where a is a constant and $2B$ converges in distribution to a χ_1^2 random variable as $n \to \infty$. Thus, the convergence rate that is stated in equation (7) is tight under the regularity conditions assumed.

2.2. Data generated from the alternative model

Bahadur and Bickel (1967) provided general conditions for which

$$\frac{1}{n}\log\{\mathrm{BF}_n(1|0)\} \xrightarrow{\mathrm{p}} c, \qquad\qquad c > 0,$$

under the alternative hypothesis for parametric tests specified according to expressions (1)–(2). Walker (2004) obtained similar results for non-parametric Bayesian models and surveyed the consistency of Bayes factors in that more general setting.

Under Walker (1969), exponential convergence of Bayes factors in favour of true alternative models can be demonstrated as follows. Assume that θ_1 denotes the data-generating value of the model parameter, and suppose that there is a $\delta > 0$ such that $N_\delta(\theta_1) \subset \Theta_1$, where $N_\delta(\theta_1) = \{\theta \in \Theta : |\theta - \theta_1| < \delta\}$. Then

$$m_0(\mathbf{X}^{(n)}) = \int_{\Theta_0} p_n(\mathbf{X}^{(n)}|\theta)\; \pi_0(\theta)\; \mathrm{d}\theta < \sup_{\theta \in \Theta_0}\{p_n(\mathbf{X}^{(n)}|\theta)\}.$$

Because

$$\lim_{n\to\infty}(P[\sup_{\theta \in \Theta - N_\delta(\theta_1)}\{L_n(\theta) - L_n(\theta_1)\} < -n\,k(\delta)]) = 1 \tag{8}$$

for some $k(\delta) > 0$ (Walker, 1969), it follows that

$$\lim_{n\to\infty}\left(P\left[\frac{m_0(\mathbf{X}^{(n)})}{\sigma_n\, p_n(\mathbf{X}^{(n)}|\theta_1)} < \exp\{-n\, k(\delta)\}\right]\right) = 1.$$

Combining this result with condition (6) (where θ_1 is now the data-generating parameter) implies that the Bayes factor in favour of the alternative hypothesis satisfies

$$\lim_{n\to\infty}\left(P\left[\frac{1}{n}\log\{\mathrm{BF}_n(1|0)\} > k(\delta)\right]\right) = 1.$$

For hypothesis tests that are conducted with local alternative prior densities, the results of this section can thus be summarized as follows.

(a) For a true null hypothesis, the Bayes factor in favour of the alternative hypothesis decreases only at rate $O_p(n^{-1/2})$.
(b) For a true alternative hypothesis, the Bayes factor in favour of the null hypothesis decreases exponentially fast.

These disparate rates of convergence imply that it is much more likely that an experiment will provide convincing evidence in favour of a true alternative hypothesis than it will for a true null hypothesis.

3. Non-local alternative prior densities

We propose two new classes of non-local prior densities which overcome the philosophical limitations that are associated with local priors and improve on the convergence rates in favour of true null hypotheses that are obtained under local alternative hypotheses. For a scalar-valued parameter θ, members of these classes satisfy the conditions $\pi(\theta) = 0$ for all $\theta \in \Theta_0$, and $\pi(\theta) > 0$ for all $\theta \in \Theta_1$. In contrast with non-local priors that are defined to be 0 on, say, an interval $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ for some $\varepsilon > 0$ (e.g. Verdinelli and Wasserman (1996) and Rousseau (2007)), the priors that we propose provide substantial flexibility in the specification of the rate at which 0 is approached at parameter values that are consistent with the null hypothesis. We can thus avoid the specification of prior densities that transition sharply between regions of the parameter space that are assigned 0 prior probability and regions that are assigned positive prior probability.

We begin by considering a test of a point null hypothesis $H_0 : \theta = \theta_0$ *versus* the composite alternative $H_1 : \theta \neq \theta_0$. In Section 3.1, we introduce the family of moment prior densities. Although this class of prior densities does not provide exponential convergence of Bayes factors in favour of true null hypotheses, it offers a substantial improvement in convergence rates over local alternative priors. Additionally, members from this class provide closed form expressions for Bayes factors in several common statistical models and yield analytic approximations to Bayes factors in large sample settings.

In Section 3.2 we introduce inverse moment prior densities, which provide exponential convergence in favour of both true null hypotheses and true alternative hypotheses. We propose extensions to vector-valued parameters in Section 3.3 and discuss default prior parameter specifications in Section 3.4.

3.1. Moment prior densities

Moment prior densities are obtained as the product of even powers of the parameter of interest and arbitrary densities. Suppose that $\pi_{\mathrm{b}}(\theta)$ denotes a base prior density with $2k$ finite integer

moments, $k \geqslant 1$, that $\pi_b(\theta)$ has two bounded derivatives in a neighbourhood containing θ_0 and that $\pi_b(\theta_0) > 0$. Then the kth *moment prior density* is defined as

$$\pi_M(\theta) = \frac{(\theta - \theta_0)^{2k}}{\tau_k} \pi_b(\theta), \tag{9}$$

where

$$\tau_k = \int_\Theta (\theta - \theta_0)^{2k} \, \pi_b(\theta) \, d\theta.$$

To simplify the exposition, we define $\tau = \tau_1$ for the first-moment prior densities. Tests of a one-sided hypothesis can be performed by taking $\pi_M(\theta) = 0$ for either $\theta < 0$ or $\theta > 0$. Note that $\pi_M(\theta_0) = 0$.

The convergence rates of Bayes factors in favour of true null hypotheses when the alternative model is specified by using a moment prior can be obtained by using Laplace approximations (Tierney *et al.*, 1989; de Bruijn, 1981). Under the null hypothesis, the maximum *a posteriori* estimate of θ, say $\tilde{\theta}_n$, satisfies $|\tilde{\theta}_n - \theta_0| = O_p(n^{-1/2})$. If the null hypothesis is true, this implies that the Bayes factor satisfies the equation (see Appendix B)

$$\mathrm{BF}_n(1|0) = O_p(n^{-k-1/2}).$$

The choice of the base prior that is used to define the moment prior can be tailored to reflect prior beliefs regarding the tails of the tested parameter under the alternative hypothesis. This choice also determines the large sample properties of Bayes factors based on the resulting moment prior when the alternative hypothesis is true. In particular, the tail behaviour of the base prior determines the finite sample consistency properties of the resulting Bayes factors. We return to this point in Section 4.

3.2. *Inverse moment priors*

Inverse moment prior densities are defined according to

$$\pi_I(\theta) = \frac{k\tau^{\nu/2}}{\Gamma(\nu/2k)} \{(\theta - \theta_0)^2\}^{-(\nu+1)/2} \exp\left[-\left\{ \frac{(\theta - \theta_0)^2}{\tau} \right\}^{-k} \right] \tag{10}$$

for $k, \nu, \tau > 0$. They have functional forms that are related to inverse gamma density functions, which means that their behaviour near θ_0 is similar to the behaviour of an inverse gamma density near 0.

Laplace approximations (Tierney *et al.*, 1989; de Bruijn, 1981) can be used with moment prior densities to obtain probabilistic bounds of the order of the marginal density of the data under the alternative hypothesis when the null hypothesis is true. Because the maximum *a posteriori* estimate satisfies

$$|\tilde{\theta}_n - \theta_0| = O_p(n^{-1/(2k+2)}),$$

it follows from Laplace's method (see Appendix B) that

$$\frac{\log\{\mathrm{BF}_n(1|0)\}}{n^{k/(k+1)}} \xrightarrow{\mathrm{p}} c \qquad \text{for some } c < 0. \tag{11}$$

For large k, inverse moment priors thus provide approximately linear convergence in n for the logarithm of the Bayes factors under both null and alternative hypotheses. In practice, values of k in the range 1–2 often provide adequate convergence rates under the null hypothesis. The

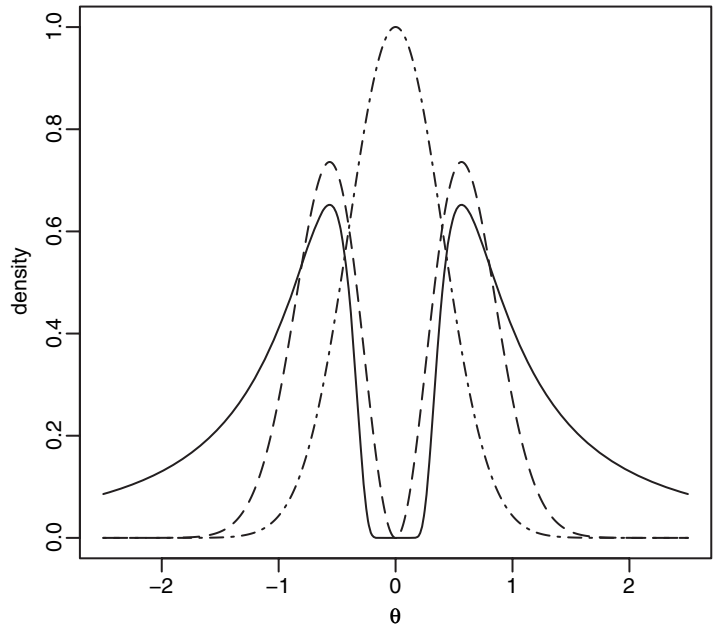

Fig. 1. Comparison of the normal moment (– – –, $k = 1$) and inverse moment (———, $k = 1$) prior densities: the moment prior approaches 0 more slowly at the origin and has lighter tails than the inverse moment prior; the Gaussian base density (· – · – ·) is depicted for reference; the moment and inverse moment priors that are depicted are used in Section 5.1 to test the value of a normal mean

selection of values of $\nu \approx 1$ implies Cauchy-like tails that may improve the power of tests under the alternative hypotheses in small sample settings.

The differences between the convergence rates of Bayes factors based on moment and inverse moment prior densities can perhaps be best understood graphically. Fig. 1 depicts an inverse moment prior density ($k = 1$) alongside a comparably scaled moment prior density ($k = 1$). The moment prior was obtained by taking $\pi_b(\theta)$ to be a normal density, which is also depicted. As is apparent from Fig. 1, the moment prior does not approach 0 as quickly near the origin as does the inverse moment prior, and so it is perhaps not surprising that these priors do not lead to exponential convergence of Bayes factors in favour of (true) point null hypotheses.

3.3. Multivariate generalizations

Both the moment and the inverse moment priors can be extended in a natural way to the multivariate setting by defining

$$Q(\boldsymbol{\theta}) = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{n\tau\sigma^2}, \tag{12}$$

where $\boldsymbol{\theta}$ is a $d \times 1$ dimensional real vector, Σ is a positive definite matrix and $\tau > 0$ is a scalar. To facilitate the exposition of results for linear models and asymptotic approximations which follow, we have incorporated some redundancy in this parameterization. In particular, the factor $1/n$ is included so that Σ^{-1}/n provides a standardized scaling when Σ^{-1} is taken to be an information matrix; the factor $1/\sigma^2$ allows us to account for the observational variance when specifying a prior on the regression coefficients in linear models and may represent an overdispersion parameter in generalized linear models. As in the univariate case, τ represents a

parameter that determines the dispersion of the prior around $\boldsymbol{\theta}_0$. In this parameterization, the inverse moment prior on $\boldsymbol{\theta}$ may be expressed as

$$\pi_{\mathrm{I}}(\boldsymbol{\theta}) = c_{\mathrm{I}}\, Q(\boldsymbol{\theta})^{-(\nu+d)/2} \exp\{-Q(\boldsymbol{\theta})^{-k}\}, \tag{13}$$

where

$$c_{\mathrm{I}} = \left|\frac{\Sigma^{-1}}{n\tau\sigma^2}\right|^{1/2} \frac{k}{\Gamma(\nu/2k)} \frac{\Gamma(d/2)}{\pi^{d/2}}. \tag{14}$$

As $Q(\boldsymbol{\theta})$ increases, the influence of the exponential term in equation (13) disappears and the tails of π_{I} become similar to those of a multivariate t-density with ν degrees of freedom. If the conditions that were specified in Walker (1969) apply, then convergence of $\log\{\mathrm{BF}_n(0|1)\}$ in favour of a true simple hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against a multivariate inverse moment prior ($k > 1$) also occurs according to expression (11) (see Appendix B).

The extension for moment priors proceeds similarly. Let $\pi_{\mathrm{b}}(\boldsymbol{\theta})$ denote a prior density on $\boldsymbol{\theta}$ for which $E_\pi[Q(\boldsymbol{\theta})^k]$ is finite. Assume also that $\pi_{\mathrm{b}}(\boldsymbol{\theta})$ has two bounded partial derivatives in a neighbourhood containing $\boldsymbol{\theta}_0$ and that $\pi_{\mathrm{b}}(\boldsymbol{\theta}_0) > 0$. Then we define the multivariate moment prior to be

$$\pi_{\mathrm{M}}(\boldsymbol{\theta}) = \frac{Q(\boldsymbol{\theta})^k}{E_{\pi_{\mathrm{b}}}[Q(\boldsymbol{\theta})^k]} \pi_{\mathrm{b}}(\boldsymbol{\theta}). \tag{15}$$

Under the preceding regularity conditions, the multivariate moment prior leads to a Bayes factor in favour of the alternative hypothesis that is $O_p(n^{-k-d/2})$ when the null hypothesis is true. Taking $k = 0$ yields the corresponding convergence rate that would be obtained by using the base prior to define the (local) alternative hypothesis.

If $\pi_{\mathrm{b}}(\boldsymbol{\theta})$ is a multivariate Gaussian density (i.e. $N_d(\boldsymbol{\theta}_0, n\tau\sigma^2\Sigma)$), then

$$E_\pi[Q(\boldsymbol{\theta})^k] = \prod_{i=0}^{k-1} (d + 2i),$$

the kth moment of a χ^2-distribution with d degrees of freedom.

If $\pi_{\mathrm{b}}(\boldsymbol{\theta})$ is a multivariate t-density with $\nu > 2$ degrees of freedom (i.e. $T_\nu(\boldsymbol{\theta}_0, n\tau\sigma^2\Sigma)$) and $k = 1$, then $E_\pi[Q(\boldsymbol{\theta})^k] = \nu d/(\nu - 2)$.

3.4. Default prior specifications

Multivariate normal moment priors contain three scalar hyperparameters—k, τ and σ, whereas multivariate t moment and inverse moment priors include an additional parameter ν. Both classes also require specification of a scale matrix Σ. In the absence of subjective prior information, we recommend two approaches for selecting default values for these hyperparameters.

In many applications, it is convenient to set Σ^{-1}/σ^2 equal to the Fisher information matrix. Scaling this matrix by n (which is implicit to the preceding parameterization of the moment and inverse moment priors) facilitates the specification of the remaining model parameters in terms of standardized effect sizes. In this parameterization, σ^2 is either assumed to represent the observational variance or is assigned a value of 1 and subsumed into the matrix Σ.

Taking $k = 1$ provides a convenient default value for both the moment and the inverse moment priors. In the former case, setting $k = 1$ and using a Gaussian base prior yields simple expressions for Bayes factors in linear models, as well as approximations to Bayes factors in large sample settings. For inverse moment priors, a value of $k = 1$ provides acceptable convergence of Bayes

factors under both the null and the alternative hypotheses without producing an unusually sharp spike at the mode of the prior density.

For the multivariate t moment prior we recommend setting $k=1$ and $\nu=3$, which produces a prior that has Cauchy-like tails. For the inverse moment prior, we recommend $\nu=1$, which again makes the tails of the prior density similar to those of a multivariate Cauchy distribution. This choice is consistent with the tails of the prior densities that were advocated by, for example, Jeffreys (1998) and more recently Bayarri and Garcia-Donato (2007). A value of $\nu=2$ may also be considered when it is convenient to have an analytic expression for the distribution function.

Finally, we recommend two methods for setting τ. In the first, we set τ so that the mode of the prior density occurs at values that are deemed most likely under the alternative hypothesis. Letting

$$w=\left(\frac{\boldsymbol{\theta}-\boldsymbol{\theta}_0}{\sigma}\right)' \frac{\Sigma^{-1}}{n} \left(\frac{\boldsymbol{\theta}-\boldsymbol{\theta}_0}{\sigma}\right),$$

the Gaussian moment prior mode occurs at the set of points for which $w=2k\tau$. The t moment prior mode occurs at points for which

$$w=\tau\frac{2\nu}{\nu-2+d},$$

and the inverse moment prior mode occurs at values of $\boldsymbol{\theta}$ for which

$$w=\tau\left(\frac{2k}{\nu+d}\right)^{1/2k}.$$

By specifying the expected differences (or standardized differences) of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}_0$, these expressions can be used to determine a value of τ that places the mode of the density at a normed distance from $\boldsymbol{\theta}_0$. This approach is useful in early stage clinical trials and other experimental settings, and can be implemented by fixing the prior modes at the response rates that are used in sample size calculations.

Alternatively, τ can be determined so that high prior probability is assigned to standardized effect sizes that are greater than a specified threshold. For instance, standardized effect sizes of less than 0.2 are often not considered substantively important in the social sciences (e.g. Cohen (1992)). This fact suggests that τ may be determined so that the prior probability that is assigned to the event that a standardized effect size is less than 0.2 is, say, less than 0.05. In the case where $k=1$, the probability that is assigned to the interval $(-a,a)$ by a scalar Gaussian moment prior centred on 0 with scale τ and $n=1$ is

$$2\left\{\Phi\left(\frac{a}{\sqrt{\tau}}\right)-\frac{a}{\sqrt{(2\pi\tau)}}\exp\left(-\frac{a^2}{2\tau}\right)-\frac{1}{2}\right\}, \tag{16}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. The corresponding probability for an inverse moment prior is

$$1-G\left\{\left(\frac{a}{\sqrt{\tau}}\right)^{-2k};\frac{\nu}{2k},1\right\}, \tag{17}$$

where $G(\cdot;c,d)$ denotes a gamma distribution function with shape c and scale d. Setting $\tau=0.114$ and $\tau=0.348$, the probabilities that are assigned by the moment prior to the interval $(-0.2,0.2)$ are 0.05 and 0.01 respectively. Similarly, for $\tau=0.133$ and $\tau=0.077$, the inverse moment prior

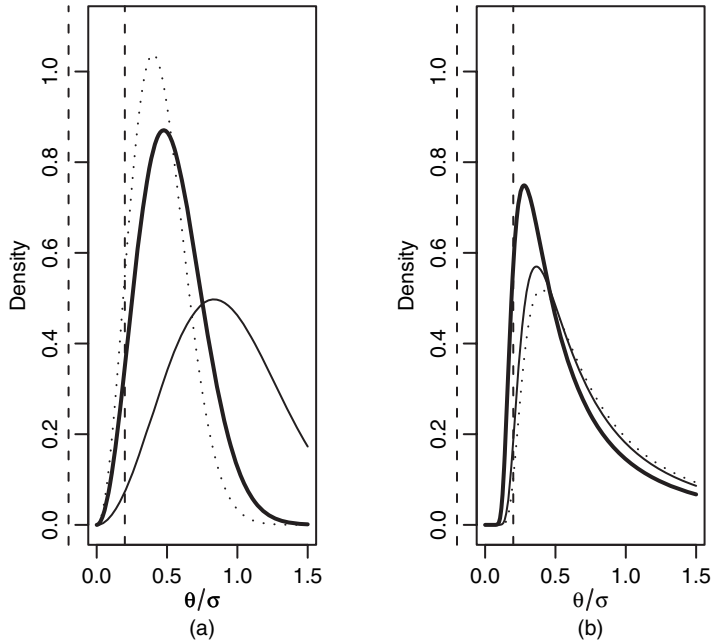

Fig. 2. Depiction of standardized priors—comparison of the (symmetric) moment and inverse moment densities on the positive θ-axis (⋮, vertical line at 0.2): (a) Gaussian moment prior densities that assign 0.01 and 0.05 prior probability to values in the interval $(-0.2, 0.2)$ (·····, moment prior that has its mode at 0.4); (b) inverse moment prior densities corresponding to (a)

assigns probabilities of 0.05 and 0.01 to the same interval. Fig. 2 depicts the standardized moment and inverse moment priors for these values. This approach for setting τ is illustrated in Section 5.4.

4. Bayes factors for linear models

In Sections 4.1 and 4.2 we address the computation of Bayes factors for linear models under moment and inverse moment priors. In Section 4.3 we discuss the extension of these results to obtain asymptotic approximations to Bayes factors in regular statistical models.

We assume that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2 I)$, where $\mathbf{y}' = (y_1, \ldots, y_n)$ is the vector of dependent variables, $\mathbf{X}$ is the design matrix, $\boldsymbol{\theta}$ is the regression coefficient vector and σ^2 is the observational variance. Letting $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ denote a partition of the parameter vector, we focus on tests of a null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_2$ is unconstrained. We let d_1 and $d_2 = d - d_1$ denote the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively.

We define $\boldsymbol{\Sigma}$ to be the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$ that corresponds to $\boldsymbol{\theta}_1$. Other definitions of $\boldsymbol{\Sigma}$ (and hence Q) might also be considered, but this choice simplifies the computation of the resulting Bayes factors and seems natural in the absence of subjective prior information regarding the departures of $\boldsymbol{\theta}_1$ from $\boldsymbol{\theta}_0$. For known σ^2, we assume a locally uniform prior for $\boldsymbol{\theta}_2$ (i.e. $\pi(\boldsymbol{\theta}_2) \propto 1$) under both hypotheses and use the usual limiting argument to cancel the constant of proportionality from the resulting Bayes factor (e.g. Berger and Pericchi (2001)). Similarly, when σ^2 is unknown we assume *a priori* that $\pi(\boldsymbol{\theta}_2, \sigma^2) \propto 1/\sigma^2$. Appendix C contains details regarding the derivation of the formulae that are provided in Sections 4.1 and 4.2.

4.1. Moment Bayes factors

We consider moment alternative prior densities that are defined by using multivariate Gaussian and t-density base measures. In the first case, a default class of alternative hypotheses may be defined by taking the base prior measure in equation (15) to be

$$\pi_{\mathrm{b}}(\boldsymbol{\theta}_1) = N(\boldsymbol{\theta}_1; \boldsymbol{\theta}_0, n\tau\sigma^2\Sigma).$$

For known σ^2, the resulting Bayes factor in favour of H_1 can be expressed as

$$\frac{\mu_k}{\prod_{i=0}^{k-1}(d_1+2i)} \frac{1}{(1+n\tau)^k} \frac{1}{(1+n\tau)^{d_1/2}} \exp\left\{ \frac{1}{2}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)'\Sigma^{-1} \frac{n\tau}{\sigma^2(1+n\tau)}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) \right\}, \tag{18}$$

where $\hat{\boldsymbol{\theta}}_1$ is the usual least squares estimate and μ_k is the kth moment of a χ^2-distribution with d_1 degrees of freedom and non-centrality parameter

$$\lambda = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)'\Sigma^{-1} \frac{n\tau}{\sigma^2(1+n\tau)}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0),$$

i.e.

$$\mu_k = 2^{k-1}(k-1)!(d_1 + k\lambda) + \sum_{j=1}^{k-1} \frac{(k-1)! \times 2^{j-1}}{(k-j)!}(d_1 + j\lambda)\mu_{k-j}. \tag{19}$$

The second term in expression (18) is the Bayes factor that is obtained under Zellner's g-prior (Zellner, 1986), which is $O_p(n^{-d/2})$ when H_0 is true. We can view the first term in expression (18) as an acceleration factor that is $O_p(n^{-k})$ and that makes the Bayes factor in favour of the alternative hypothesis $O_p(n^{-k-d_1/2})$ when the null hypothesis is true. For example, setting $k = d_1/2$ doubles the rate at which evidence in favour of H_0 is accumulated.

Now consider the case in which σ^2 is unknown. Although there is no simple expression for an arbitrary value of k, for $k = 1$ the Bayes factor is

$$\frac{d_1 + \hat{\lambda}}{d_1(1+n\tau)}(1+n\tau)^{(n-d)/2} \left\{ 1 + n\tau \frac{s_{\mathrm{R}}^2(n-d)}{s_{\mathrm{R}}^2(n-d) + (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)'\Sigma^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)} \right\}^{-(n-d_2)/2}, \tag{20}$$

where $s_{\mathrm{R}}^2(n-d)$ is the sum of squared residuals,

$$\hat{\lambda} = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)'\Sigma^{-1} \frac{n\tau}{\hat{\sigma}^2(1+n\tau)}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)$$

and

$$\hat{\sigma}^2 = \frac{s_{\mathrm{R}}^2(n-d)}{n-d_2} + \frac{(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)'\Sigma^{-1}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)}{(1+n\tau)(n-d_2)}. \tag{21}$$

As $n \to \infty$, $\hat{\sigma}^2$ converges in probability to σ^2 and $d_1 + \hat{\lambda}$ converges to $\mu_1 = d_1 + \lambda$ as defined in equation (19). Bayes factors based on normal moment priors with $k > 1$ may be derived along similar lines but have substantially more complicated expressions.

As noted by Liang *et al.* (2008), Bayes factors that are based on Zellner's g-prior do not have finite sample consistency for the preceding test, i.e. this Bayes factor violates the 'information paradox' because it does not become unbounded as the least squares estimate $\hat{\boldsymbol{\theta}} \to \infty$ for a fixed sample size. The Bayes factor in expression (20) inherits this property from its g-prior base density, and for a fixed sample size n is bounded by

$$(p_1 + n\tau)(n - p_2)(1+n\tau)^{(n-p)/2-1}/p_1,$$

which is $O(n^{(n-p)/2+1})$. For comparison, the corresponding bound on the Bayes factor that is obtained by using the g-prior base density is $(1+n\tau)^{(n-p)/2-1}$, which is $O(n^{(n-p)/2})$. We do not regard the lack of finite sample consistency that is exhibited in expression (20) to be of practical concern, and we note that strong evidence can be obtained in favour of the alternative hypothesis in the test of a scalar parameter θ for $n \geqslant 4$ when $\tau = 1$. However, the Bayes factor in expression (20) can be modified so that it has finite sample consistency either by using empirical Bayes methods to set τ, or through the specification of a hyperprior density on τ. Details concerning both of these methods are provided in Liang *et al.* (2008).

Bayes factors that are defined from multivariate t base densities can be obtained by expressing the base density as a scale mixture of normals, i.e., by exploiting the fact that a multivariate t-density on ν degrees of freedom can be expressed as a scale mixture of normals, it follows that the t-moment Bayes factor can be expressed as

$$\int_0^\infty \mathrm{BF_N}(g\tau)\frac{g(\nu-2)}{\nu}\,\mathrm{IG}\left(g;\frac{\nu}{2},\frac{\nu}{2}\right)\mathrm{d}g, \tag{22}$$

where $\mathrm{IG}(g;\nu/2,\nu/2)$ denotes an inverse gamma probability density function and $\mathrm{BF_N}(g\tau)$ is the Bayes factor that is obtained under a normal moment prior with $k=1$ after substituting $g\tau$ for τ in either expression (18) (σ^2 known) or expression (20) (σ^2 unknown). A similar procedure was suggested by Liang *et al.* (2008) for obtaining the Zellner–Siow Cauchy prior from a scale mixture of Zellner's g-prior. For known σ^2, the resulting Bayes factor has finite sample consistency. For the case of unknown σ^2, the resulting Bayes factor has finite sample consistency whenever $\nu < n-d+2$. In particular, finite sample consistency is achieved for $\nu = 3$ whenever $n \geqslant d+2$.

4.2. *Inverse moment Bayes factors*

Bayes factors for linear models are not available in closed form for inverse moment alternative prior densities. However, they can be calculated by evaluating either a simple univariate integral or a bivariate integral, depending on whether or not σ^2 is assumed to be known *a priori*.

For known σ^2, the Bayes factor in favour of the alternative hypothesis is

$$\left(\frac{2}{n\tau}\right)^{d_1/2}\frac{k\,\Gamma(d_1/2)}{\Gamma(\nu/2k)}\frac{E_z[(n\tau/z)^{(\nu+d_1)/2}\exp\{-(n\tau/z)^k\}]}{\exp(-\frac{1}{2}\lambda_n)}, \tag{23}$$

where z denotes a non-central χ^2 random variable with d_1 degrees of freedom and non-centrality parameter

$$\lambda_n = (\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\frac{\Sigma^{-1}}{\sigma^2}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0).$$

When σ^2 is not known *a priori*, the default Bayes factor in favour of the alternative hypothesis is

$$\left(\frac{2}{n\tau}\right)^{d_1/2}\frac{k\,\Gamma(d_1/2)}{\Gamma(\nu/2k)}\frac{E_{(w,z)}[(n\tau/z)^{(\nu+d_1)/2}\exp\{-(n\tau/z)^k\}]}{[1+(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\{\Sigma^{-1}/s_{\mathrm{R}}^2(n-d)\}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)]^{-(n-d_2)/2}}. \tag{24}$$

Here, the expectation is taken with respect to the joint distribution of the random vector (w,z), where w is distributed as an inverse gamma random variable with parameters $(n-d_2)/2$ and $s_{\mathrm{R}}^2(n-d)/2$, and z given w has as a non-central χ^2-distribution on d_1 degrees of freedom and non-centrality parameter

$$\lambda_n = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)' \frac{\Sigma^{-1}}{w} (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0).$$

For $\nu = 1$, inverse moment Bayes factors exhibit finite sample consistency properties that are similar to Bayes factors based on the Zellner–Siow prior and t-moment priors with $\nu = 3$.

4.3. Asymptotic approximations to default Bayes factors

The default Bayes factors that were obtained in the previous subsection for linear models can be combined with the methodology that was presented in Section 3 to obtain approximations to Bayes factors in large sample settings.

Suppose that $x_1, \ldots, x_n$ denote independent draws from a distribution that has density function $f(x|\boldsymbol{\theta}, \sigma^2)$, $\boldsymbol{\theta} \in \mathcal{R}^d$, $\sigma^2 \in \mathcal{R}^+$. Suppose further that the parameter vector $\boldsymbol{\theta}$ is partitioned into components $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, where $\boldsymbol{\theta}_1$ is the parameter of interest, $\boldsymbol{\theta}_2$ is a nuisance parameter and σ^2 is a dispersion parameter (set to $\sigma^2 = 1$ for no overdispersion). Consider the test of a point null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$, $\boldsymbol{\theta}_0 \in \mathcal{R}^{d_1}$.

Under Walker's (1969) regularity conditions, if the prior density on $\boldsymbol{\theta}$ is continuous and positive in a neighbourhood of the true parameter value, then the posterior distribution of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ converges to a multivariate normal distribution with mean equal to the maximum likelihood estimate $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ and asymptotic covariance matrix $\hat{\sigma}^2 \hat{V}$, where $\hat{\sigma}^2$ is a consistent estimate of σ^2 and $\hat{V}$ is the inverse of the observed information matrix. If Σ denotes the submatrix of $\hat{V}$ corresponding to $\boldsymbol{\theta}_1$, then moment and inverse moment alternative priors can be used to define default alternative hypotheses. Taking $\sigma^2 = \hat{\sigma}^2$, the Bayes factor that is obtained under the moment alternative prior specification can be approximated by expression (18); the Bayes factor in favour of the alternative hypothesis using an inverse moment prior can be approximated by expression (23).

5. Examples

5.1. Test of a normal mean

To contrast the performance of local and non-local alternative priors, let $X_1, \ldots, X_n$ denote independent and indentically distributed (IID) $N(\theta, 1)$ data, and consider a test of $H_0 : \theta = 0$ against each of the following alternative hypotheses:

$$\begin{aligned} H_1^{\mathrm{a}} &: \pi(\theta) = N(\theta; 0, 2), \\ H_1^{\mathrm{b}} &: \pi(\theta) = \mathrm{Cauchy}(\theta), \\ H_1^{\mathrm{c}} &: \pi(\theta) \propto (\theta^2)^{-1} \exp(-0.318/\theta^2), \\ H_1^{\mathrm{d}} &: \pi(\theta) \propto \theta^2 \, n(\theta; 0, 0.159), \end{aligned}$$

where Cauchy$(\cdot)$ refers to a standard Cauchy density. The non-local densities that define H_1^{c} and H_1^{d} are depicted in Fig. 1. Hypothesis H_1^{a} corresponds to an intrinsic prior for testing the null hypothesis H_0 (Berger and Pericchi, 1996b), whereas hypothesis H_1^{b} corresponds to Jeffreys's recommendation for testing H_0 (Jeffreys, 1998). The parameters of the inverse moment prior in H_1^{c} were $\upsilon = 1, k = 1$ and $\tau = 1/\pi = 0.318$, which means that the inverse moment prior's tails match those of the Cauchy prior in H_1^{b}. The modes of the resulting inverse moment prior occurred at $\pm 1/\sqrt{\pi} = \pm 0.564$. To facilitate comparisons between the moment and inverse moment priors, we chose the parameters of the normal moment prior so that the two densities had the same modes.

As this set-up is a particular case of linear models with a known variance, the Bayes factors for H_1^{c} and H_1^{d} can be computed from expressions (23) and (18) respectively. The Bayes factor in

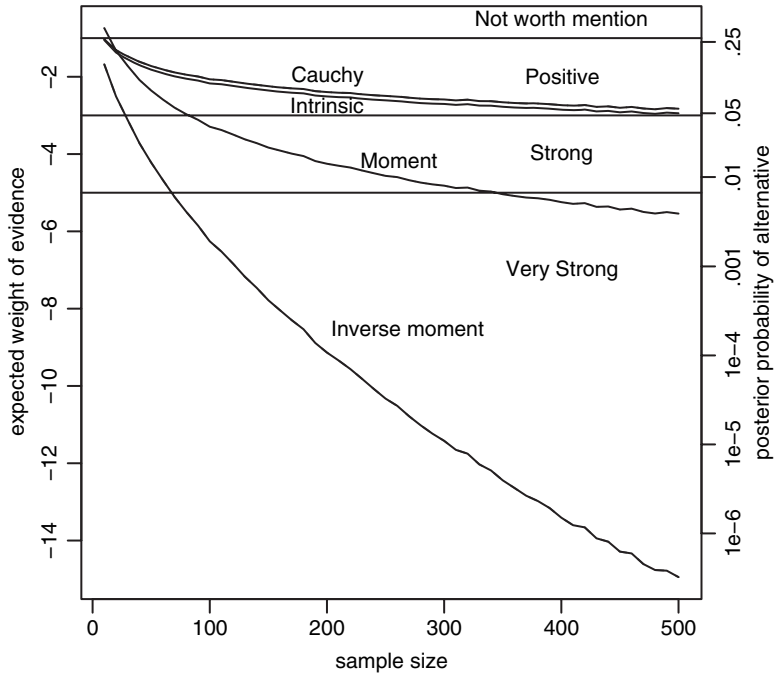

Fig. 3. Expected weight of evidence under four alternative hypotheses for the test of a normal mean (Section 5.1): the expected value of the weight of evidence was computed for IID $N(0,1)$ data as a function of sample size, which is depicted on the horizontal axis; the vertical axis on the right-hand side of the plot indicates the posterior probability of the alternative hypothesis corresponding to the expected weights of evidence that are depicted by the four curves under the assumption that the null and alternative hypotheses are assigned equal probability *a priori*; Kass and Raftery's (1995) categorization of weight of evidence ('not worth more than a bare mention', 'positive evidence', 'strong evidence' or 'very strong evidence') is provided for reference

favour of H_1^a can be expressed as $\exp(0.5n^2\bar{x}^2/a_n)/\sqrt{(\tau a_n)}$, where $a_n = n + 1/\tau$ and $\tau = 2$. The Bayes factor that is associated with H_1^b can be evaluated numerically as a one-dimensional integral.

The performance of the non-local alternative priors *versus* the local alternative prior is illustrated in Fig. 3 for data that were simulated under the null model. Each curve represents an estimate of the expected 'weight of evidence' (i.e. logarithm of the Bayes factor) based on 4000 simulated data sets of the sample size indicated. As Fig. 3 demonstrates, the local alternative hypotheses are unable to provide what Kass and Raftery (1995) termed strong support in favour of the null hypothesis even for sample sizes exceeding 500. In contrast, strong support is achieved, on average, for sample sizes of approximately 30 and 85 under H_1^c and H_1^d respectively, whereas very strong evidence is obtained for sample sizes of 70 and 350. The local Bayes factors require more than 30000 observations, on average, to achieve very strong evidence in favour of a true null hypothesis.

When null and alternative hypotheses are assigned equal probability *a priori*, we note that strong evidence against a hypothesis implies that its posterior probability is less than 0.047, whereas very strong evidence against a hypothesis implies that its posterior probability is less than 0.0067.

It is important to note that non-local alternative hypotheses often provide stronger evidence against the null hypothesis when the null hypothesis is false than do local alternative hypotheses. Heuristically, non-local alternative priors take mass from parameter values that are

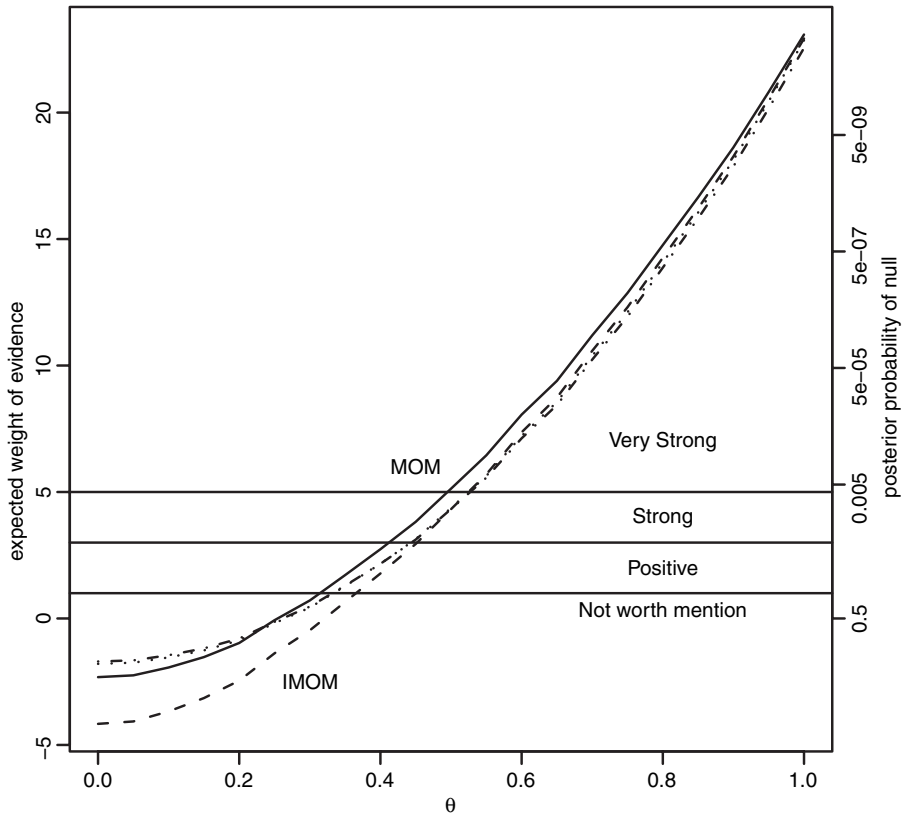

Fig. 4. Expected weight of evidence under four alternative hypotheses for the test of a normal mean (Section 5.1): the expected value of the weight of evidence was computed for IID $N(\theta, 1)$ data for a fixed sample size of 50; the vertical axis on the right-hand side of the plot indicates the posterior probability of the null hypothesis corresponding to the expected weights of evidence depicted by the four curves under the assumption that the null and alternative hypotheses were assigned equal probability *a priori*; the weights of evidence obtained by using moment and inverse moment priors are depicted as full and broken curves respectively; the weights of evidence that are provided by the intrinsic and Cauchy priors nearly overlap and are slightly larger than the moment and inverse moment curves at the origin; Kass and Raftery's (1995) categorization of weight of evidence is provided for reference

close to Θ_0 and reassign the mass to parameter values that are considered likely under the alternative hypothesis. This reassignment increases the expected weight of evidence that is collected in favour of both true null and true alternative hypotheses. This fact is illustrated in Fig. 4, where the expected weight of evidence is depicted as a function of a known value of θ and a fixed sample size of 50. For values of $0.25 < \theta < 1.05$, the moment prior provides, on average, stronger weight of evidence in favour of the alternative hypotheses than either of the local alternatives, whereas the inverse moment prior provides essentially equal or stronger weight evidence than the local alternatives for all values of $\theta > 0.5$.

Such gains do not, of course, accrue without some trade-off. For the moment and inverse moment priors that were selected for this example and a sample size of 50, the expected weight of evidence in favour of the null hypothesis is at least positive for all values of $\theta < 0.28$ when the alternative hypothesis is based on the inverse moment prior, or when $\theta < 0.19$ and the alternative is based on the moment prior. Indeed, for values of $\theta < 0.16$, the expected weight of

evidence in favour of the null hypothesis falls in Kass and Raftery's (1995) strong range for the inverse-moment-based alternative. In contrast, the local priors provide positive expected weight of evidence in favour of the null hypothesis only for values of $\theta < 0.17$.

More generally, local alternative priors can be expected to provide greater evidence in favour of true alternative hypotheses whenever the data-generating parameter is closer to the null value than it is to regions of the parameter space that are assigned non-negligible probability under a non-local alternative. Of course, obtaining positive evidence in favour of an alternative hypothesis for parameter values that are very close to Θ_0 requires large sample sizes, regardless of whether a local or non-local prior is used to define the alternative hypothesis. Such extensive data are often not collected unless the investigator suspects *a priori* that θ is close to (but not in) Θ_0, which in turn would result in the specification of a non-local alternative model that concentrated its mass more closely around Θ_0.

We note that the weight of evidence that is obtained in favour of the alternative hypothesis increases rapidly for large values of $|\theta|$ under all the alternative hypotheses. For large θ-values, the tails of the prior that define the alternative hypothesis are largely irrelevant, provided only that the prior has not been taken to be so disperse that the Lindley paradox applies.

5.2. Test of a normal variance

We next consider the application of non-local priors to the calculation of Bayes factors for the test of a normal variance. Let $X_1, \ldots, X_n$ denote IID $N(0, \zeta)$ observations, and define the null hypothesis to be $H_0 : \zeta = 1$. We consider the following three alternative hypotheses:

$$H_1^{\mathrm{a}} : \pi(\zeta) = \frac{1}{\pi\sqrt{\zeta(1+\zeta)}};$$

$$H_1^{\mathrm{b}} : \pi(\zeta) = c(\zeta - \zeta_0)^2 \zeta^{-\alpha-1} \exp\left(-\frac{\lambda}{\zeta}\right),$$

$$c = \frac{\lambda^\alpha}{\lambda^2\,\Gamma(\alpha-2) - 2\zeta_0\lambda\,\Gamma(\alpha-1) + \zeta_0^2\,\Gamma(\alpha)}, \qquad \zeta_0 = 1, \alpha = 5, \lambda = 4;$$

$$H_1^{\mathrm{c}} : \pi(\zeta) = \frac{\alpha}{\pi\zeta\,\log(\zeta)^2} \exp\left\{-\frac{\alpha}{\log(\zeta)^2}\right\}, \qquad \alpha = 0.2.$$

The first alternative hypothesis represents an intrinsic prior (Bertolino *et al.*, 2000) for the test of H_0. The second alternative hypothesis is a moment prior based on an inverse gamma density. The parameters of this density were chosen so that it has modes at 0.5 and 2. The third alternative hypothesis was obtained from an inverse moment prior for $\log(\zeta)$ centred on 0 with parameters $\tau = 0.25$, $k = 1$ and $\nu = 1$. On the logarithmic scale, this density has modes at ± 0.5. The three densities that define the alternative hypotheses are depicted in Fig. 5.

Fig. 6 illustrates the relationship between the average weight of evidence (i.e. the logarithm of the Bayes factor) in favour of the three alternative hypotheses as a function of sample size n for data that are generated under the null hypothesis. For each value of n, the average weight of evidence in favour of the alternative hypothesis was estimated by generating 4000 random samples of size n under the null hypothesis that $\zeta = 1$.

The trend that is depicted in Fig. 6 for the intrinsic Bayes factor (top curve) clearly illustrates the poor performance of Bayes factors that are based on local priors in this setting. On average, in excess of 350 observations are needed to obtain strong evidence (Kass and Raftery, 1995) in favour of the null hypothesis when the intrinsic prior is used to define an alternative hypothesis,

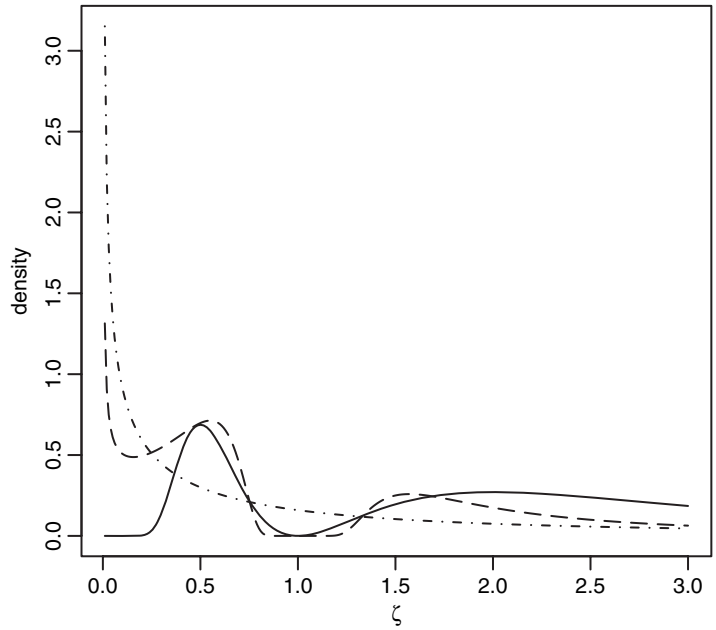

Fig. 5. Prior densities on variance parameter ζ used for the hypothesis tests that are described in Section 5.2: the intrinsic prior ($\cdot - \cdot - \cdot$) is smoothly decreasing; the inverse moment prior ($- \; - \; -$) is flatter than the moment prior (———) at $\zeta = 1$, where both densities are 0

and very strong evidence is not typically obtained with even 10000 observations. In contrast, the average weight of evidence in favour of the null hypothesis is strong with fewer than 100 observations by using either of the non-local alternatives, and the average weight of evidence becomes very strong with only 375 and 200 observations by using the moment and transformed inverse moment priors respectively.

We note that the expected weight of evidence in favour of alternative hypotheses that is obtained by using the non-local alternatives is similar to the intrinsic prior when $\zeta \neq 1$. For example, Fig. 7 illustrates the expected weight of evidence that is obtained under H_1^a–H_1^c as a function of ζ based on a sample size of $n = 100$.

5.3. Stylized Bayesian clinical trial

To illustrate the effect of non-local priors in a common hypothesis testing problem, consider a highly stylized phase II trial of a new drug that is aimed at improving the overall response rate from 20% to 40% for some population of patients with a common disease. Such trials are not designed to provide conclusive evidence in favour of regulatory approval of a drug but instead attempt to provide preliminary evidence of efficacy. The null hypothesis of no improvement is formalized by assuming that the true overall response rate of the drug is $\theta = 0.2$. For illustration, we consider three specifications of the alternative hypothesis:

$$H_1^a : \pi(\theta) \propto \theta^{-0.8}(1-\theta)^{-0.2} \qquad 0.2 < \theta < 1;$$
$$H_1^b : \pi(\theta) \propto \theta^{0.2}(1-\theta)^{0.8} \qquad 0.2 < \theta < 1;$$
$$H_1^c : \pi(\theta) \propto (\theta - 0.2)^{-3} \exp\{-0.055/(\theta - 0.2)^2\} \qquad 0.2 < \theta < 1.$$

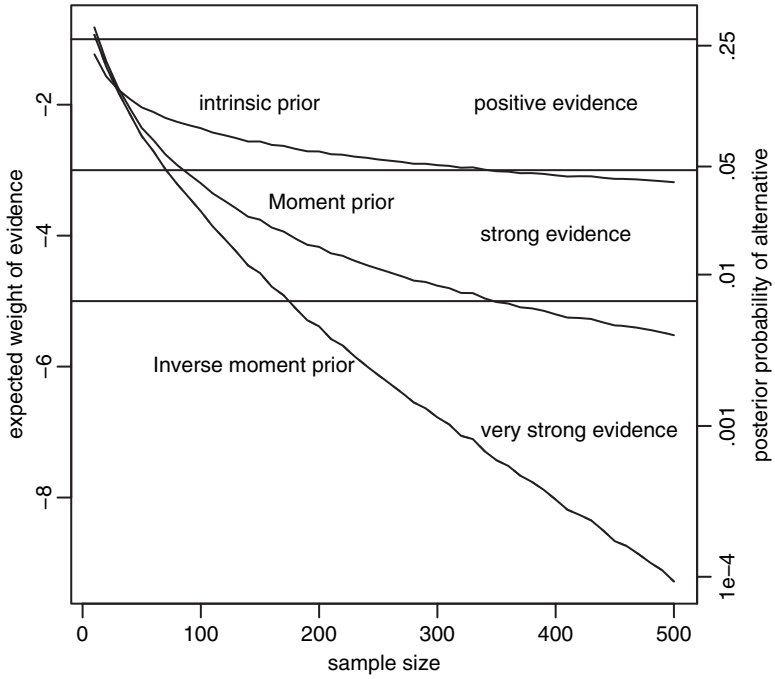

Fig. 6. Expected weight of evidence under three alternative hypotheses for the test of a normal variance (Section 5.2): the expected value of the weight of evidence was computed for IID $N(0, 1)$ data as a function of sample size, which is depicted on the horizontal axis; the vertical axis on the right-hand side of the plot indicates the posterior probability of the alternative hypothesis corresponding to the expected weights of evidence depicted by the three curves under the assumption that the null and alternative hypotheses were assigned equal probability *a priori*; Kass and Raftery's (1995) categorization of weight of evidence is provided for reference

In each case, the prior densities are assigned a value of 0 outside the interval (0.2,1) since $\theta > 0.2$ under the alternative hypotheses.

Hypothesis H_1^a represents the truncation of a local alternative hypothesis centred on $\theta = 0.2$ and can be expected to approximate the behaviour of one-sided tests that are conducted using intrinsic priors, intrinsic Bayes factors or the non-informative prior. Hypothesis H_1^b represents a mildly informative prior centred on the target value for the response rate of the new drug. The final hypothesis, H_1^c, denotes the inverse moment prior truncated to the interval (0.2,1) having parameters $\nu = 2$, $\tau = 0.055$ and $k = 1$.

To complete the specification of the trial, suppose that patients are accrued and the trial is continued until one of two events occurs:

(a) the posterior probability that is assigned to either the null or alternative hypothesis exceeds 0.9, or
(b) 50 patients have entered the trial.

Trials that are not stopped before the 51st patient accrues are assumed to be inconclusive. Assume further that the null and alternative hypotheses are assigned equal weight *a priori*, and that the null hypothesis is actually true. Using these assumptions, in Table 1 we list the proportions of trials that end with a conclusion in favour of each hypothesis, along with the average number of patients observed before each trial is stopped.

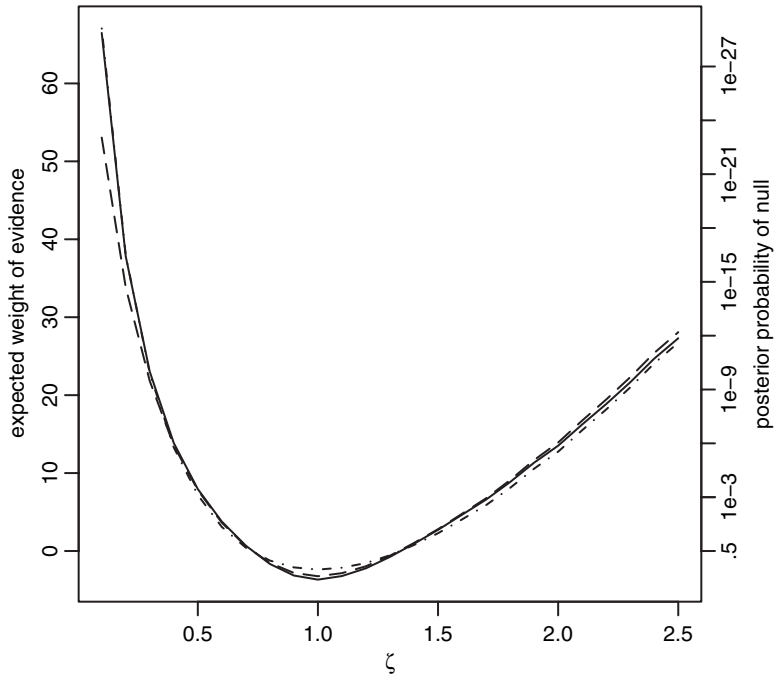

Fig. 7. Expected weight of evidence under three alternative hypotheses for the test of a normal variance (Section 5.2): the expected value of the weight of evidence was computed for IID $N(0, \theta)$ data for a fixed sample size of 100; the expected weight of evidence was computed under three alternative hypotheses based on the intrinsic ($\cdot - \cdot - \cdot$), moment ($- \; - \; -$) and inverse moment (———) priors as a function of the variance parameter ζ; for $\zeta \ll 1$, the hypothesis based on the moment prior provides less power than the hypotheses based on either intrinsic or inverse moment densities; for $\zeta \approx 2.5$, the moment prior provides slightly more evidence in favour of the alternative than does the inverse moment prior, which in turn provides slightly more evidence than the intrinsic prior

Table 1. Operating characteristics of the hypothetical clinical trial in Section 5.3

Hypothesis	*Proportion of trials stopped for* H_0	*Proportion of trials stopped for* H_1^*	*Average number of patients enrolled*
H_1^a	0.43	0.11	35.3
H_1^b	0.51	0.05	37.1
H_1^c	0.91	0.04	17.7

The probability of stopping a trial early in favour of the null hypothesis was approximately twice as great for the inverse moment alternative as it was for either of the truncated beta priors. In addition, only half as many patients were required to stop the trial when the inverse moment prior was used. In considering these figures, it is important to note that the parameters for the truncated beta densities were chosen so that all tests had approximately the same power when the alternative hypothesis of $\theta = 0.4$ was true. For $\theta = 0.4$, the powers of the three hypotheses (estimated from 1000 simulated trials) were 0.825, 0.823 and 0.812 respectively. Thus, by balancing the rates at which evidence accumulated under the null and alternative models, the probability of correctly accepting the null hypothesis was significantly increased when the null

hypothesis was true, whereas the probability of accepting the alternative hypothesis when it was true was not adversely affected at the targeted value of θ.

5.4. Approximate Bayes factors for probit regression

In our final example, we performed a simulation study to illustrate the asymptotic approximations to Bayes factors based on non-local alternative priors in the context of simple binary regression models. For each sample size n, we generated y_i, $i = 1, \ldots, n$, as Benoulli random variables with probit success probabilities $\Phi(\theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i})$, where $\theta_1 = 0.7$, $\theta_0 = \theta_2 = 0$ and $(x_{1,i}, x_{2,i})$ were simulated as mean 0, variance 1, normal random variables with correlation 0.5. 1000 data sets with sample sizes of $n = 50, 100, 200$ were used to test two pairs of hypotheses:

$$\begin{aligned} H_{0,1} &: \theta_1 = 0 \quad \textit{versus} \quad H_{1,1} : \theta_1 \sim \pi_1(\theta), \\ H_{0,2} &: \theta_2 = 0 \quad \textit{versus} \quad H_{1,2} : \theta_2 \sim \pi_2(\theta). \end{aligned} \tag{25}$$

The prior densities that were assumed under the alternative hypotheses were a normal moment prior, a Cauchy inverse moment prior (i.e. $\nu = 1$) and Zellner's g-prior with the default value $\tau = 1$. This value of τ produced posterior probabilities in favour of including the first covariate in the regression model that were similar to the probabilities that obtained by using the non-local priors and thus provides a useful contrast for examining the convergence of the posterior probabilities in favour of a true null hypothesis. We tested the same three specifications of τ for the non-local alternative priors that were used in Section 5.3.

In Table 2, we show the average posterior probabilities in favour of each alternative hypothesis when the null and alternative models were assigned equal probability *a priori*. As expected, the non-local alternative hypotheses provide substantially more evidence in favour of the true null hypothesis H_{02} that $\theta_2 = 0$ than does Zellner's g-prior. Even the average posterior probability of $H_{1,1}$ is approximately the same under all prior assumptions. For example, the 1% inverse moment prior assigned essentially the same average posterior probability to $H_{1,1}$ as did Zellner's g-prior, but only between 60% to 17% as much probability to $H_{1,2}$ as n was increased from 50 to 200.

Table 2. Expected posterior probabilities for the inclusion of covariates in the simulated probit regression model data of Section 5.4

	Probabilities for the following priors:						
	Moment			*Inverse moment*			*Zellner's g-prior*
	1%	*5%*	*0.4*	*1%*	*5%*	*0.4*	
n = 50							
$P(\theta_1 \neq 0 \mid \text{data})$	0.67	0.80	0.83	0.76	0.79	0.74	0.76
$P(\theta_2 \neq 0 \mid \text{data})$	0.06	0.17	0.21	0.12	0.17	0.10	0.20
n = 100							
$P(\theta_1 \neq 0 \mid \text{data})$	0.92	0.96	0.97	0.95	0.96	0.95	0.95
$P(\theta_2 \neq 0 \mid \text{data})$	0.03	0.09	0.12	0.06	0.10	0.04	0.15
n = 200							
$P(\theta_1 \neq 0 \mid \text{data})$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$P(\theta_2 \neq 0 \mid \text{data})$	0.01	0.05	0.07	0.02	0.05	0.02	0.12

6. Discussion

This paper has focused on the description of two classes of non-local alternative prior densities that can be used to define Bayesian hypothesis tests that have favourable sampling properties, and the comparison of their sampling properties to tests defined by using standard local alternative priors. To a large extent, we have ignored philosophical issues regarding the logical necessity to specify an alternative hypothesis that is distinct from the null hypothesis. In general, it is our view that one hypothesis (and a test statistic) is enough to obtain a p-value, but that two hypotheses are required to obtain a Bayes factor. The two classes of prior densities that were proposed in this paper provide functional forms that may be used to specify the second hypothesis when detailed elicitation of a subjective prior is not practical. Members of both classes eliminate the high costs that are associated with standard Bayesian methods when the null hypothesis is true. They also allow specification of a parameter τ that implicitly defines what is meant by a substantively meaningful difference between the null and alternative hypotheses. Not only does the specification of this parameter aid in the interpretation of test findings, but also judicious choice of its value can increase the evidence that is reported in favour of both true null and true alternative hypotheses.

In many applications in the social sciences, standardized effect sizes of approximately 0.2 are regarded as substantively important. When this is so, the three choices of τ that were illustrated in Sections 5.3 and 5.4 for moment and inverse moment priors may provide useful default prior specifications under the alternative hypothesis. The simulation results that were provided in Section 5.4 may be used as a guide to select from among these choices.

The density functions that were proposed in this paper may be extended for use as alternatives to composite null hypotheses in various ways. For example, if the null hypothesis is defined according to $H_0 : \theta \sim U(a, b)$, then an inverse moment prior might be specified according to

$$\pi_{\mathrm{I}}(\theta) = \begin{cases} \dfrac{k\tau^{\nu/2}}{\Gamma(\nu/2k)} \{(\theta-a)^2\}^{-(\nu+1)/2} \exp\left[-\left\{ \dfrac{(\theta-a)^2}{\tau} \right\}^{-k} \right] & \theta < a, \\ \dfrac{k\tau^{\nu/2}}{\Gamma(\nu/2k)} \{(\theta-b)^2\}^{-(\nu+1)/2} \exp\left[-\left\{ \dfrac{(\theta-b)^2}{\tau} \right\}^{-k} \right] & \theta > b. \end{cases} \tag{26}$$

Other forms of moment and inverse moment priors may also be specified for vector-valued parameter vectors. Alternatively, prior distributions that decrease to 0 near the boundaries between disjoint null and alternative parameter spaces might be considered.

In addition to standard hypothesis tests, non-local alternative prior densities may also provide important advantages over local alternative priors in model selection algorithms. In that setting, the assignment of high probability to 'true' null hypotheses of no effect may eliminate or reduce the need to specify strong prior penalties against the inclusion of any model covariate.

Much of the methodology that is described in this paper has been implemented in the R package `mombf` (Rossell, 2008).

Acknowledgements

We thank James Berger, Luis Pericchi, Lee Ann Chastain and several referees for numerous comments that improved the presentation of material in this paper.

Appendix A: Regularity conditions

It is assumed that the following conditions from Walker (1969) hold. To facilitate comparison, we have retained Walker's numbering system for these conditions.

Assumption A1. Θ is a closed set of points in R^s.

Assumption A2. The sample space $\mathcal{X} = \{x : f(x|\theta) > 0\}$ is independent of θ.

Assumption A3. If θ_1 and θ_2 are distinct points of Θ, the set

$$\{x : f(x|\theta_1) \neq f(x|\theta_2)\} > 0$$

has Lebesgue measure greater than 0.

Assumption A4. Let $x \in \mathcal{X}$ and $\theta' \in \Theta$. Then for all θ such that $|\theta - \theta'| < \delta$, with δ sufficiently small,

$$|\log\{f(x|\theta)\} - \log\{f(x|\theta')\}| < H_\delta(x, \theta'),$$

where

$$\lim_{\delta \to 0}\{H_\delta(x, \theta')\} = 0,$$

and, for any $\theta_0 \in \Theta$,

$$\lim_{\delta \to 0}\left\{\int_{\mathcal{X}} H_\delta(x, \theta')\ f(x|\theta_0)\ \mathrm{d}\mu\right\} = 0.$$

Assumption A5. If Θ is not bounded, then for any $\theta_0 \in \Theta$, and sufficiently large Δ,

$$\log\{f(x|\theta)\} - \log\{f(x|\theta_0)\} < K_\Delta(x, \theta_0)$$

whenever $|\theta| > \Delta$, where

$$\lim_{\delta \to 0}\left\{\int_{\mathcal{X}} K_\Delta(x, \theta_0)\ f(x|\theta_0)\ \mathrm{d}\mu\right\} < 0.$$

For the remaining conditions, let θ_0 be an interior point of Θ.

Assumption B1. $\log\{f(x|\theta)\}$ is twice differentiable with respect to θ in some neighbourhood of θ_0.

Assumption B2. The matrix $\mathbf{J}(\boldsymbol{\theta}_0)$ with elements

$$J_{ij}(\boldsymbol{\theta}_0) = \int_{\mathcal{X}} f_0 \frac{\partial \log(f_0)}{\partial \theta_{0,i}} \frac{\partial \log(f_0)}{\partial \theta_{0,j}}\ \mathrm{d}\mu,$$

where f_0 denotes $f(x|\boldsymbol{\theta}_0)$, is finite and positive definite. In the scalar case, this condition becomes $0 < J(\theta_0) < \infty$, where

$$J(\theta_0) = \int_{\mathcal{X}} f_0 \left\{\frac{\partial \log(f_0)}{\partial \theta_0}\right\}^2 \mathrm{d}\mu.$$

Assumption B3.

$$\int_{\mathcal{X}} \frac{\partial f_{0,i}}{\partial \theta_{0,i}}\ \mathrm{d}\mu = \int_{\mathcal{X}} \frac{\partial^2 f_0}{\partial \theta_{0,i}\ \partial \theta_{0,j}}\ \mathrm{d}\mu = 0.$$

Assumption B4. If $|\theta - \theta_0| < \delta$, where δ is sufficiently small, then

$$\left|\frac{\partial^2 \log\{f(x|\theta)\}}{\partial \theta_i\ \partial \theta_j} - \frac{\partial^2 \log\{f(x|\theta_0)\}}{\partial \theta_{0,i}\ \partial \theta_{0,j}}\right| < M_\delta(x, \theta_0),$$

where

$$\lim_{\delta \to 0}\left\{\int_{\mathcal{X}} M_\delta(x, \theta_0) f(x|\theta_0)\ \mathrm{d}\mu\right\} = 0.$$

Appendix B

B.1. Convergence of Bayes factors based on inverse moment priors

Assume that the regularity conditions of Walker (1969) hold, and that $\theta_0 = 0$ is the data-generating parameter. Without loss of generality, we take $\tau = 1$. Then the marginal density of the data under the alternative hypothesis can be expressed as

$$\begin{aligned} m_1(\mathbf{X}^{(n)}) &= c \int_\Theta \exp\left\{ -(\theta^2)^{-k} + L_n(\theta) - \frac{(\nu+1)}{2} \log(\theta^2) \right\} \mathrm{d}\theta \\ &\equiv c \int_\Theta \exp\{h(\theta)\} \, \mathrm{d}\theta \end{aligned}$$

for a constant c that is independent of n. Expanding $L_n(\theta)$ around a maximum likelihood estimate $\hat{\theta}_n$ and differentiating $h(\theta)$ implies that the values of θ that maximize $h(\theta)$, say $\tilde{\theta}_n$, satisfy

$$2k\tilde{\theta}_n(\tilde{\theta}_n^2)^{-k-1} + L_n''(\theta^*)(\tilde{\theta}_n - \hat{\theta}_n) - \frac{\nu+1}{\tilde{\theta}_n} = 0 \tag{27}$$

for some $\theta^* \in (\tilde{\theta}_n, \hat{\theta}_n)$. (If more than one maximum likelihood estimate exists, the difference between them is $o_p(n^{-1/2})$.) Rearranging terms leads to

$$n(\tilde{\theta}_n^2)^{k+1} \left(1 - \frac{\hat{\theta}_n}{\tilde{\theta}_n} \right) = \frac{2k - (\nu+1)(\tilde{\theta}_n^2)^k}{-L_n''(\theta^*)/n}. \tag{28}$$

From regularity condition B4, the denominator on the right-hand side of equation (28) converges in probability to $J(\theta_0)$, whereas the first term in the numerator is constant. Noting that $\hat{\theta}_n = O_p(n^{-1/2})$, it follows that the two maxima of the posterior satisfy

$$\operatorname{plim}(n^r \tilde{\theta}_n) = \pm \frac{(2k)^r}{J(\theta_0)^r}, \tag{29}$$

where $r = 1/(2k+2)$.

By expanding around each of the maxima, Walker's (1969) conditions allow Laplace's method to be used to obtain an asymptotic approximation to $m_1(\mathbf{X}^{(n)})$, yielding

$$m_1(\mathbf{X}^{(n)}) \approx c \left\{ \frac{(4k^2+2k)}{\tilde{\theta}_n}^{2(k+1)} - L_n''(\tilde{\theta}_n) \right\}^{-1/2} |\tilde{\theta}_n|^{-\nu-1} \exp\{-(\tilde{\theta}_n^2)^{-k} + L_n(\tilde{\theta}_n)\}. \tag{30}$$

Expanding the log-likelihood function around $\hat{\theta}_n$ yields

$$L_n(\tilde{\theta}_n) = L_n(\hat{\theta}_n) + \tfrac{1}{2} L_n''(\theta^*)(\tilde{\theta}_n - \hat{\theta}_n)^2$$

for some $\theta^* \in (\tilde{\theta}_n, \hat{\theta}_n)$. When the null hypothesis is true, $L_n(\hat{\theta}_n) - L_n(\theta_0) = O_p(1)$, and it follows that

$$\operatorname*{plim}_{n\to\infty} [n^s \log\{\mathrm{BF}_n(1|0)\}] = -\left\{ \frac{2k}{J(\theta_0)} \right\}^s - \frac{(2k)^{2r}}{2J(\theta_0)^{2r-1}}$$

where $s = -k/(k+1)$.

An extension to the multivariate setting can be accomplished by using similar arguments. Let s_{ij} denote the (i, j)th element of the positive definite matrix $\mathbf{S} = \Sigma^{-1}/\sigma^2$, and suppose that $\boldsymbol{\theta}_0 = \mathbf{0}$. Then the values of $\boldsymbol{\theta}$ that provide the local maxima of the marginal density of the data under the alternative model satisfy the system of equations

$$0 = -\frac{(\nu+d)\sum_j \theta_j s_{ij}}{Q(\boldsymbol{\theta})} + 2k\, Q(\boldsymbol{\theta})^{-k-1} \sum_j \theta_j s_{ij} + \sum_j \frac{\partial L_n(\boldsymbol{\theta}^*)}{\partial \theta_i \theta_j} (\theta_j - \hat{\theta}_j) \tag{31}$$

for $i = 1, \ldots, d$ and $\boldsymbol{\theta}^* \in (\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. Rearranging terms implies that each maximum of $h(\boldsymbol{\theta})$, say $\boldsymbol{\theta}^{(m)}$, satisfies

$$\operatorname*{plim}_{n\to\infty}\left\{ n\ Q(\tilde{\boldsymbol{\theta}}_n^{(m)})^{k+1} - \frac{\sum_j J_{ij}\tilde{\theta}_{n,j}^{(m)}}{2k\sum_j s_{ij}\tilde{\theta}_{nj}^{(m)}}\right\} = 0, \qquad i = 1,\ldots,d.$$

Because $\mathbf{S}$ is positive definite, it follows that $\operatorname{plim}\{n^{1/(k+1)}\ Q(\tilde{\boldsymbol{\theta}}_n^{(m)})\} = c^{(m)}$ for some $c^{(m)} > 0$. Thus $\operatorname{plim}(n^{1/(2k+2)}\tilde{\theta}_{n,j}) = d_j^{(m)}$, for some vector $\mathbf{d}^{(m)}$ for which $\max_j |d_j^{(m)}| > 0$. Expanding $L_n(\tilde{\boldsymbol{\theta}}_n)$ around the maximum likelihood estimate gives

$$L_n(\tilde{\boldsymbol{\theta}}_n) = L_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{2}\sum_{i,j}\frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\theta_i\partial\theta_j}(\tilde{\theta}_{n,i}^{(m)} - \hat{\theta}_{n,i})(\tilde{\theta}_{n,i}^{(m)} - \hat{\theta}_{n,i})$$

for some $\boldsymbol{\theta}^* \in (\hat{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\theta}}^{(m)})$. The second term on the right-hand side of this equation satisfies

$$\operatorname*{plim}_{n\to\infty}\left\{\frac{1}{2}n^s\sum_{i,j}\frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\theta_i\,\partial\theta_j}(\tilde{\theta}_{n,i}^{(m)} - \hat{\theta}_{n,i})(\tilde{\theta}_{n,i}^{(m)} - \hat{\theta}_{n,i})\right\} = b^{(m)}$$

for some $b^{(m)} < 0$. Applying Laplace's method at each maximum and summing, it follows that

$$\lim_{n\to\infty}(P[n^s\log\{\mathrm{BF}_n(1|0)\} < a]) = 1, \qquad \text{for some } a < 0.$$

B.2. Convergence of Bayes factors based on moment priors

Assume again that Walker's (1969) conditions hold, and that $\theta_0 = 0$ is the data-generating parameter. Then the marginal density of the data under the alternative hypothesis can be expressed as

$$\begin{aligned} m_1(\mathbf{X}^{(n)}) &= \frac{1}{\tau_k}\int_\Theta \exp[2k\log(\theta) + \log\{\pi(\theta)\} + L_n(\theta)]\,\mathrm{d}\theta \\ &\equiv \frac{1}{\tau_k}\int_\Theta \exp\{h(\theta)\}\,\mathrm{d}\theta. \end{aligned}$$

The value of θ that maximizes $h(\theta)$, $\tilde{\theta}_n$, satisfies

$$\frac{2k}{\tilde{\theta}_n} + \frac{\pi'(\tilde{\theta}_n)}{\pi(\tilde{\theta}_n)} + L_n'(\tilde{\theta}_n) = 0.$$

Expanding the derivative of the log-likelihood function around $\hat{\theta}_n$ leads to

$$\frac{2k}{n} + \frac{\tilde{\theta}_n\,\pi'(\tilde{\theta}_n)}{n\,\pi(\tilde{\theta}_n)} + \frac{1}{n}L_n''(\theta^*)\tilde{\theta}_n(\tilde{\theta}_n - \hat{\theta}_n) = 0$$

for some $\theta^* \in (\tilde{\theta}_n, \hat{\theta}_n)$, from which it follows that $|\tilde{\theta}_n - \hat{\theta}_n| = O_p(n^{-1})$.

The Laplace approximation to $m_1(\mathbf{X}^{(n)})$ around $\tilde{\theta}_n$ is

$$m_1(\mathbf{X}^{(n)}) \approx \frac{\sqrt{(2\pi)}}{\tau_k}\left\{\frac{2k}{\tilde{\theta}_n^2} - L_n''(\tilde{\theta}_n)\right\}^{-1/2}\exp[2k\log(\tilde{\theta}_n) + \log\{\pi(\tilde{\theta}_n)\} + L_n(\tilde{\theta}_n)]. \tag{32}$$

Because $|L_n(\tilde{\theta}_n) - L_n(\theta_0)| = O_p(1)$, the Bayes factor in favour of the alternative hypothesis when the null hypothesis is true thus satisfies

$$\mathrm{BF}_n(1|0) = O_p(n^{-k-1/2}). \tag{33}$$

An extension to the multivariate moment prior densities is straightforward. As in the scalar case, the maximum *a posteriori* estimate satisfies $|\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n| = O_p(n^{-1})$. Application of the Laplace approximation to the d-dimensional integral defining the marginal density of the data under the alternative hypothesis implies that

$$\mathrm{BF}_n(1|0) = O_p(n^{-k-d/2}). \tag{34}$$

Appendix C: Bayes factors for linear models

Without loss of generality, we assume that X_1 is orthogonal to X_2, i.e. $X_1'X_2 = \mathbf{0}$, and we assume that the notation and assumptions of Section 4 apply.

To compute Bayes factors between models, it suffices to evaluate the marginal density of the sufficient statistic under each hypothesis. For linear models with a known σ^2, the least squares estimator $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ is a sufficient statistic under both H_0 and H_1. From the orthogonality of $\mathbf{X}_1$ and $\mathbf{X}_2$, the Bayes factor can be computed as

$$\frac{\int N(\hat{\boldsymbol{\theta}}_1; \boldsymbol{\theta}_1, \sigma^2\Sigma)\, \pi(\boldsymbol{\theta}_1|\sigma^2)\, \mathrm{d}\boldsymbol{\theta}_1}{N(\hat{\boldsymbol{\theta}}_1; \boldsymbol{\theta}_0, \sigma^2\Sigma)}, \tag{35}$$

where $N(\cdot; \mathbf{m}, V)$ is the multivariate normal density with mean m and covariance V, Σ is the submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\boldsymbol{\theta}_1$ and $\pi(\boldsymbol{\theta}_1|\sigma^2)$ is the prior distribution for $\boldsymbol{\theta}_1$. For an unknown σ^2, the sufficient statistic is completed by adding s_{R}^2, the usual unbiased estimate for σ^2, and the Bayes factor can be expressed as

$$\frac{\int \left\{ \int N(\hat{\boldsymbol{\theta}}_1; \boldsymbol{\theta}_1, \sigma^2\Sigma)\, \pi(\boldsymbol{\theta}_1|\sigma^2)\, \mathrm{d}\boldsymbol{\theta}_1 \right\} G\{s_{\mathrm{R}}^2; (n-p)/2, (n-p)/2\sigma^2\}(1/\sigma)\, \mathrm{d}\sigma^2}{\int N(\hat{\boldsymbol{\theta}}_1; \boldsymbol{\theta}_0, \sigma^2\Sigma)\, G\{s_{\mathrm{R}}^2; (n-p)/2, (n-p)/2\sigma^2\}(1/\sigma)\, \mathrm{d}\sigma^2}, \tag{36}$$

where $G(\cdot; a, b)$ denotes the gamma density with shape parameter a and scale parameter b.

C.1. Moment-based Bayes factor

When $\pi(\theta_1|\sigma^2)$ takes the form of a normal moment prior density and σ^2 is known, the numerator in expression (35) can be re-expressed as

$$\frac{\exp[-\frac{1}{2}\{\hat{\boldsymbol{\theta}}_1'(\Sigma^{-1}/\sigma^2)\hat{\boldsymbol{\theta}}_1 + \boldsymbol{\theta}_0'(\Sigma^{-1}/n\tau\sigma^2)\boldsymbol{\theta}_0 - (\hat{\boldsymbol{\theta}}_1 + \boldsymbol{\theta}_0/n\tau)'(\Sigma^{-1}/\sigma^2)n\tau/(n\tau+1)(\hat{\boldsymbol{\theta}}_1 + \boldsymbol{\theta}_0/n\tau)\}]}{\left\{ \prod_{i=0}^{k-1}(d_1+2i) \right\} (2\pi\sigma^2)^{d_1/2}|\Sigma|^{1/2}(1+n\tau)^{d_1/2+k}}$$
$$\times \int \{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)'V^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\}^k N(\boldsymbol{\theta}_1; \mathbf{m}, V)\, \mathrm{d}\boldsymbol{\theta}_1, \tag{37}$$

where

$$\mathbf{m} = \frac{n\tau}{n\tau+1}\hat{\boldsymbol{\theta}}_1 + \frac{1}{n\tau+1}\boldsymbol{\theta}_0$$

and

$$V^{-1} = \frac{1+n\tau}{n\tau\sigma^2}\Sigma^{-1}.$$

The second term in expression (37) represents the kth non-central moment of a normal distribution. Denoting the Cholesky decomposition of V by C, it follows that $C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \sim N\{C(\mathbf{m} - \boldsymbol{\theta}_0), I\}$. Hence, $(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)'V^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)$ follows a χ^2-distribution with d_1 degrees of freedom and non-centrality parameter

$$\lambda = (\mathbf{m} - \boldsymbol{\theta}_0)'V^{-1}(\mathbf{m} - \boldsymbol{\theta}_0) = \left(\hat{\boldsymbol{\theta}}_1 - \frac{n\tau-1}{n\tau}\boldsymbol{\theta}_0\right)' \Sigma^{-1} \frac{n\tau}{\sigma^2(1+n\tau)} \left(\hat{\boldsymbol{\theta}}_1 - \frac{n\tau-1}{n\tau}\boldsymbol{\theta}_0\right).$$

Simple algebra then shows that expression (35) equals expression (18).

For an unknown σ^2, we consider only the case where $k = 1$. Noting that the inner integral is equal to expression (37) and noting the similarity of the outer integral to an inverse gamma distribution, the numerator in expression (36) can be expressed as

$$\frac{\Gamma\{(n-d_2)/2\}}{2\{(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\Sigma^{-1}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)/2(1+n\tau)+s_{\mathrm{R}}^2(n-p)/2\}^{(n-d_2)/2}} \times\left\{d_1+\frac{\hat{\lambda}\hat{\sigma}^2}{(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\Sigma^{-1}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)/(1+n\tau)(n-d_2)+s_{\mathrm{R}}^2(n-p)/(n-d_2)}\right\}, \tag{38}$$

where

$$\hat{\lambda}=(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\Sigma^{-1}\frac{n\tau}{\hat{\sigma}^2(1+n\tau)}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)$$

and $\hat{\sigma}^2$ is as in equation (21). On simplifying the denominator in equation (36), it follows that expression (36) is equal to expression (20).

Computing t-moment Bayes factors is straightforward, using that the multivariate T can be expressed as a mixture of multivariate normals with respect to an inverse gamma variance parameter g.

C.2. Inverse-moment-based Bayes factor

Setting π equal to π_{I} in equation (13), the numerator in expression (35) can be written as

$$c_{\mathrm{I}}\int\left[\left(\frac{n\tau}{z}\right)^{(\nu+d_1)/2}\exp\left\{-\left(\frac{n\tau}{z}\right)^k\right\}\right]N(\boldsymbol{\theta}_1;\hat{\boldsymbol{\theta}}_1,\sigma^2\Sigma)\,\mathrm{d}\boldsymbol{\theta}_1, \tag{39}$$

where

$$z=(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)'\frac{\Sigma^{-1}}{\sigma^2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0).$$

This integral represents the expected value of $(n\tau/z)^{(\nu+d_1)/2}\exp\{-(n\tau/z)^k\}$ when θ_1 arises from an $N(\hat{\boldsymbol{\theta}}_1,\sigma^2\Sigma)$ distribution or, equivalently, when z arises from a χ^2-distribution with d_1 degrees of freedom and non-centrality parameter

$$\lambda_n=(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\frac{\Sigma^{-1}}{\sigma^2}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0).$$

It follows that expression (35) is equal to expression (23).

For an unknown σ^2, similar derivations show that the numerator in expression (36) is equal to

$$\left|\frac{\Sigma^{-1}}{n\tau}\right|^{1/2}\frac{k}{\Gamma(\nu/2k)}\frac{\Gamma(d_1/2)}{\pi^{d_1/2}}\left(\frac{n-p}{2}\right)^{-d_1/2}(s_{\mathrm{R}}^2)^{-d_1/2-1}\frac{\Gamma\{(n-d_2)/2\}}{2\Gamma\{(n-p)/2\}} \times E_{w,z}\left[\left(\frac{n\tau}{z}\right)^{(\nu+d_1)/2}\exp\left\{-\left(\frac{n\tau}{z}\right)^k\right\}\right], \tag{40}$$

where z, given w, has a χ^2-distribution with d_1 degrees of freedom and non-centrality parameter

$$\lambda_n=(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0)'\frac{\Sigma^{-1}}{w}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0),$$

and w is marginally distributed as an inverse gamma distribution with parameters $(n-d_2)/2$ and $s_{\mathrm{R}}^2(n-p)/2$. After some manipulation, it follows that expression (36) is equal to expression (24).

References

Bahadur, R. R. and Bickel, P. J. (1967) Asymptotic optimality of Bayes' test statistics. *Technical Report*. University of Chicago, Chicago. Unpublished.

Bayarri, M. J. and Garcia-Donato, G. (2007) Extending conventional prior for testing general hypotheses in linear models. *Biometrika*, **94**, 135–152.

Berger, J. O. and Mortera, J. (1999) Default Bayes factors for nonnested hypothesis testing. *J. Am. Statist. Ass.*, **94**, 542–554.

Berger, J. O. and Pericchi, L. R. (1996a) The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Ass.*, **91**, 109–122.
Berger, J. O. and Pericchi, L. R. (1996b) The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 23–42. Oxford: Oxford University Press.
Berger, J. O. and Pericchi, L. R. (1998) Accurate and stable Bayesian model selection: the median intrinsic Bayes factor. *Sankhya* B, **60**, 1–18.
Berger, J. O. and Pericchi, L. R. (2001) Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection* (ed. P. Lahiri), pp. 135–193. Fountain Hills: Institute of Mathematical Statistics Press.
Bertolino, F., Racugno, W. and Moreno, E. (2000) Bayesian model selection approach to analysis of variance under heteroscedasticity. *Statistician*, **49**, 503–517.
de Bruijn, N. G. (1981) *Asymptotic Methods in Analysis*. New York: Dover Publications.
Cano, J. A., Kessler, M. and Moreno, E. (2004) On intrinsic priors for nonnested models. *Test*, **13**, 445–463.
Casella, G. and Moreno, E. (2006) Objective Bayesian variable selection. *J. Am. Statist. Ass.*, **101**, 157–167.
Cohen, J. (1992) A power primer. *Psychol. Bull.*, **112**, 155–159.
Conigliani, C. and O'Hagan, A. (2000) Sensitivity of the fractional Bayes factor to prior distributions. *Can. J. Statist.*, **28**, 343–352.
De Santis, F. and Spezzaferri, F. (2001) Consistent fractional Bayes factor for nested normal linear models. *J. Statist. Planng Inf.*, **97**, 305–321.
Jeffreys, H. (1998) *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.
Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.
Lahiri, P. (ed.) (2001) *Model Selection*. Beachwood: Institute of Mathematical Statistics.
Levine, R. A. and Casella, G. (1996) Convergence of posterior odds. *J. Statist. Planng Inf.*, **55**, 331–344.
Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008) Mixtures of g priors for Bayesian variable selection. *J. Am. Statist. Ass.*, **103**, 410–423.
Lindley, D. V. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.
Moreno, E. (2005) Objective Bayesian methods for one-sided testing. *Test*, **14**, 181–198.
Moreno, E., Bertolino, F. and Racugno, W. (1998) An intrinsic limiting procedure for model selection and hypotheses testing. *J. Am. Statist. Ass.*, **93**, 1451–1460.
Moreno, E. and Girón, F. J. (2005) Consistency of Bayes factors for intrinsic priors in normal linear models. *C. R. Math.*, **340**, 911–914.
O'Hagan, A. (1995) Fractional Bayes factors for model comparison. *J. R. Statist. Soc.* B, **57**, 99–118.
O'Hagan, A. (1997) Properties of intrinsic and fractional Bayes factors. *Test*, **6**, 101–118.
Pérez, J. M. and Berger, J. O. (2002) Expected-posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.
Rossell, D. (2008) mombf: moment and inverse moment Bayes factors. *R Package Version 1.0.3*.
Rousseau, J. (2007) Approximating interval hypothesis: p-values and Bayes factors. In *Bayesian Statistics 8* (eds J. M. Bernardo, M. J. Bayarri, J. Berger and A. P. Dawid), pp. 417–452. Oxford: Clarendon.
Tierney, L., Kass, R. and Kadane, J. B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Statist. Ass.*, **84**, 710–716.
Verdinelli, I. and Wasserman, L. (1996) Bayes factors, nuisance parameters, and imprecise tests. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 765–771. Oxford: Clarendon.
Vlachos, P. K. and Gelfand, A. E. (2003) On the calibration of Bayesian model choice criteria. *J. Statist. Planng Inf.*, **111**, 223–234.
Walker, A. M. (1969) On the asymptotic behaviour of posterior distributions. *J. R. Statist. Soc.* B, **31**, 80–88.
Walker, S. G. (2004) Modern Bayesian asymptotics. *Statist. Sci.*, **19**, 111–117.
Walker, S. G. and Hjort, N. L. (2001) On Bayesian consistency. *J. R. Statist. Soc.* B, **63**, 811–821.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distribution. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (eds P. K. Goel and A. Zellner), pp. 233–243. Amsterdam: North-Holland.